DOCUMENT RESUME

ED 200 646 TM 810 241

AUTHOR Stollie, Theodore R.; And Others
TITLE An Empirical Comparison of Title I NCE Gains Estimated with Model A1 and with Model A2.
PUB DATE Apr 79
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, CA, April, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; *Criterion Referenced Tests; Elementary Education; *Equated Scores; *Mathematical Models; *Norm Referenced Tests; *Reading Comprehension
IDENTIFIERS Chicago Criterion Referenced Record; *Chicago Public Schools IL; Iowa Tests of Basic Skills; Normal Curve Equivalent Scores; *Title I Evaluation and Reporting System

ABSTRACT

A comparison is made between Elementary Secondary Education Act Title I normal curve equivalent (NCE) gains derived from the norm-referenced test (NRT) Model A1 and Title I NCE gains derived from the criterion-referenced test (CRT) Model A2. In addition, four equating methods (normalized, curvilinear, linear, and abbreviated equipercentile) for obtaining NCE Title I gains with Model A2 are discussed, and the need for matched samples on both the NRT and CRT required for Model A2 is examined. Data consisted of the Iowa Test of Basic Skills and the Chicago Criterion Referenced Record in Word Attack and Comprehension Skills reading comprehension test results from the 1977-78 Chicago Title I evaluation of students aged 8-11. Correlations between the NRT and CRT were too low (less than .6) to justify equating them. Recommendations were made that a simulation study be designed to investigate the issues addressed in this study. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200646

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

An Empirical Comparison of Title I NCE Gains Estimated with Model A1 and with Model A2

Theodore R. Storlie, Educational Testing Service
William Rice, Chicago Board of Education
Philip Harvey, Educational Testing Service
Laura R. Crane, Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

T. R. Storlie

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

paper presented at the annual meeting of
The American Educational Research Association
San Francisco, April, 1979

TM 810 241

Introduction

RMC models for Title I evaluation include a procedure for use of criterion-referenced tests (CRT) instead of norm-referenced tests (NRT) to obtain estimates of project impact in terms of the NCE metric (Tallmadge and Wood, 1978, p.7) . The major focus of this study was to compare estimates of project impact derived under standard Model A1 procedures (using a NRT) with estimates derived under Model A2 procedures (using a CRT). Two additional substudies were also planned. The first, Substudy A was designed to compare equating methods while Substudy B was designed to examine the stability of the equating methods.

Model A2, like Model A1, has been subjected to criticism with respect to the equipercentile assumption. Model A2, however, has been severely critized on additional grounds. The model requires a minimum NRT-CRT correlation of .6 in order for the equating to be valid. Linn has indicated that this restriction "is much too lenient" (1978, p. 12). Linn did a simulated study using Anchor Test Study data and found the Model A2 expected no treatment performance was 3.6 NCE units higher than the Model A1 expected no treatment performance. His point was that "even under equating conditions much better than can generally be expected for Model A2 applications, systematic errors may be introduced simply due to the equating." (Linn, p. 14).

Kearns reported a study designed to compare all three standard RMC Models (A1, B1, and C1) with each other and with their CRT counterparts (A2, B2, C2); because of several problems with the data, the analysis was halted at the point of presenting mean NCE gains for each of the six methods, however, the Model A1 and Model A2 gains were very different for reading, -2.7 vs 8.9 for Model A1 and Model A2 respectively, while the correlation between NRT and CRT was .56 (Kearns, 1978, p.p. 6-7). Linn's study indicated Model A2 introduced a

negative bias while Kearns' study possibly indicated a positive bias.

Angoff, in addressing the issue of comparable scores, concluded "If the decisions that are made on the basis of a table of comparable scores are made in the central region of the distribution, if the tests are relatively similar in function, if they have approximately equal reliabilities and validities, and finally, if generalizations are restricted to the kind of group that formed the basis of conversion in the first place, these are the conditions that allow the free use of comparable scores." (Angoff, 1966, p 264). In the context of Title I evaluation, where essentially median scores are equated, it would seem that use of CRTs should be acceptable under certain conditions. In the same article, Angoff identifies four sources of error in equating scores "across tests of the same function... One source of error lies in the unreliability of the measuring instruments themselves, and affects the stability of statistics obtained in the collection of data. A second source of error lies in the design of the equating experiment and the method of treating the data: the various methods of equating are not equally reliable. A third source of error lies in the choice of samples used to establish the conversion line. Different samples will yield lines that differ randomly from one another. Finally, there is a source of error that is characteristic of any statistical enterprise - the error associated with the size of the study sample. Obviously, larger samples will yield more reliable data." (Angoff, 1966, p 253). Additionally, for nonparallel tests conversion lines will differ: depending on the definition of comparability (e.g. standard score deviates, equivalent percentiles, equal predicted scores on some criterion), depending on the population used, and depending on differential selection of equating groups (Angoff, 1966, pp 254-255). Some of the problems in equating nonparallel tests that Angoff addresses do not actually apply to

equating of CRTs to NRTs in the Title I evaluation Models. Angoff is addressing the issue of developing equating tables for use in the sense a norms table would be used, therefore, concerns for characteristics of the groups used to establish the equivalency tables are not relevant to this study. However, reliability, sample size, and equating methods are very relevant concerns.

The Technical Assistance Center (TAC) subcommittee on Model A2 recommended a hierarchial order of preference for equating CRTs to NRTs. The hierarchy, in order of preference, was. "a. Develop normalized test score distributions for the non-normed and normed tests at the local level... b. Use a curvilinear analogue for equating tests across the entire score range as described in Angoff (1971)... c. Use a linear analog for equating tests as described in Angoff (1971)... d. Use the current A2 strategy." (Hansen, 1978, p 24).

This study was designed to focus on three aspects of Model A2: the "equivalence" of gains estimated using Model A2 procedures vs gains using Model A1 procedures; the differences between Model A2 gain estimates derived by different equating techniques; and the stability of the various equating techniques.

Data Source

The Chicago Board of Education (CBE) administers the Iowa Tests of Basic Skills (ITBS) and the Chicago Criterion Referenced Record in Word Attack and Comprehension Skills (CRT) annually every spring as a part of its citywide testing and Title I evaluation programs. For purposes of this study, the

reading comprehension subtests of the ITBS and CRT were used.

For standardized testing, CBE employs functional level testing on a citywide basis and in addition has developed local norms for the ITBS (Rice, 1977; Wick and Ward; Storlie, Rice, Johnson and Crowe, 1979). Thus three types of scores were available for equating, raw scores, GE scores, and the Chicago based NCE score (CSS).

The CCRT was primarily designed for students between the ages of seven and eleven (although older students were frequently administered the test) and has two levels. For reading comprehension, Level I of the CCRT contains 50 items and Level II contains 81 items (40 of which were also in Level I). Reliabilities were quite high, above .90, for ages nine through eleven, and are reported in Table 1 for the sample data of the study.

CBE staff merged the FY78-ITBS Title I tape (which also contained FY77 results for FY78 participants) with the FY78-CCRT tape and the FY77-CCRT tape. From the total of 39,063 student records on the merged tape, a total of 4,004 students, aged eight through eleven in FY78, had CCRT data for both FY77 and FY78. The total of 4,004 students with pre (FY77) and post (FY78) CCRT data represented approximately one-third of the number of students with pre-and post ITBS data for ages eight through eleven. The number of student records available for use in the study was less than the 4,004 matched records because only students with the same CCRT level pre- and post were included in the study. A further inclusion criterion imposed was that only student records with pre and post ITBS data were retained. This was a deviation from study as proposed, but the number of records without pre- and post ITBS data indicated that this was a reasonable deletion. Table 1 reports the full breakdown of the original 4,004 student records.

Design of the Study

In designing the study, some technical issues discussed in literature concerned with equating techniques had to be addressed. For example, Angoff discusses use of random groups each taking one test, or use of random groups each taking one of the tests and both taking a common equating test for developing a table of comparable scores (Angoff, 1971, p. 591). By way of contrast, nothing in the RMC Users Guide (Tallmadge and Wood, 1978) addresses the issue of developing random groups, presumably because most LEAs would not have sufficient sample size and the purpose is not to develop an equating table that would be used over time. However, the design of this study, did employ Angoff's Design I procedure of random groups each "taking one test".

For the main study, the A1 vs. A2 comparison, the total sample was randomly split into two subgroups for each age level. Groups were then randomly assigned to the NRT or CRT situation for equating. The two subgroups were random groups for developing the equating table as in Angoff's Design I (1971, p. 569). To compare A1 and A2 NCE gain estimates, the NRT subgroup was then to be used to derive the A1 NCE gain and the CRT subgroup was to be used to derive the A2 gain estimates. To avoid possible treatment bias, pretest scores were to be used for developing the equating tables (Hansen, 1978, p. 25). For this study each of the four equating methods were to be applied to derive separate equating tables.

For Substudy A, comparison of the four equating methods, only age groups ten and eleven had sufficiently large n's (see Table 1) to allow for eight randomly formed subgroups, two subgroups to be randomly assigned to each equating method (one for the NRT group and one for the CRT group).

For Substudy B, stability estimates of the equating methods, four randomly generated subgroups were to be assigned at random to NRT_a, CRT_a, NRT_b, and CRT_b, where the 'a' and 'b' subscripts refer to separate equating samples for the same equating procedure. Subgroups were to be reformed for each equating procedure.

Because of functional level testing, use of ITBS raw scores for equating purposes would have led to an additional level of analysis, ITBS test level. The resulting loss of data (due to insufficient n's) led to consideration of two alternatives: equating with the ITBS/GE score, an expanded scale score and the first derived scale score for the ITBS, which with spring to spring testing does not suffer from the usual problems of a GE scale; or equating with the CBE local norm scale that was derived from the ITBS/GE scores. Since Angoff recommended using local norms when possible (Angoff, 1971, p. 596) and the Chicago based score (CSS) yields NCE gain estimates not substantially different from the RMC procedure for tests with local and national norms (Storlie, et al., 1979, p. 10); it was decided to use the CSS metric for equating the CRT scores.

The specific equating procedures (normalized, curvilinear, linear and RMC) are described below:

> Normalized Equating - This procedure requires normalizing the two distributions and then equating on the basis of normalized standard scores, i.e., raw scores from each distribution are considered equivalent if they have the same normalized standard score. Magnusson (1966, pp. 235-238) gives a procedure for normalizing scores which consists of determining cumulative frequencies for class interval means and transforming the cumulative frequencies to proportions. The proportions are then transformed to the normalized z-score by use of a special table (cf. Fisher and Yates, 1953).

Curvilinear Equating - This procedure requires plotting the relative cumulative frequency of each distribution on probability paper without normalizing and then equating corresponding percentiles (Angoff, 1971, p. 571). Both Angoff (1971, p. 571) and Wood (1978, p. 101) recommend the curves be smoothed prior to equating by plotting adjusted frequencies (analytically adjusted unless one is very experienced with hand smoothing techniques) such as are obtained by the Cureton and Tukey rolling weighted average method.

Linear Equating - This procedure requires solving the linear equation given below:

$$NRT = \left(\frac{SD_{NRT}}{SD_{CRT}}\right)CCRT + NRT_{mean} - \left(\frac{SD_{NRT}}{SD_{CRT}}\right)CCRT_{mean}$$

where SD refers to standard deviation (Angoff, 1971, p. 569).

RMC Equating (abbreviated equipercentile) - This procedure required equating the median NRT score to the median CCRT score.

Given that pretest data were to be used for equating, regardless of equating procedure, the procedure for computing NCE gains would then consist of (1) determining the no treatment expectation (NCEe) derived from pretest NRT median (converted to %ile to NCE) and (2) converting the posttest CRT median to the NCE metric by equating the posttest CRT median to the equivalent NRT pretest score and then using the NRT tables to determine a posttest %ile and corresponding observed NCE value (NCEo). NCE gains then equals NCEo - NCEe.

Results

The results of this study were very disappointing in that it was found that the NRT - CRT correlations were less than .6 for all age groups (pre- and posttest). Correlations are given in Table 2A and descriptive statistics are given in Table 2B. It was decided at this point to discontinue analysis since any

comparisons would be confounded with the lack of "overlap" between tests. Quite obviously, since the CRT reliabilities were high, the ITBS and CRT reading comprehension substests were measuring substantially different facets of reading comprehension.

Summary and Conclusions

A study was designed to examine the differences between Model A1 and Model A2 NCE gain estimates. Additional designs to examine effects of different equating techniques on Model A2 NCE gain estimates and to examine the stability of NCE gain estimates within the equating techniques applied were also planned.

Unfortunately, the correlations between the normed-referenced and criterion-referenced tests were too low to justify equating them. Since the CRT distributions were not normal (see Table 2B), the comparisons of equating methods would have been particularly useful. At this point, we would recommend a simulation study be designed to investigate the issues this paper attempted to address with empirical data.

Table 1

NUMBER OF STUDENTS (N) NORM-REFERENCED (ITBS) AND CRITERION REFERENCED (CRT) TEST DATA FOR FY77 AND FY78.

	Age 8 N	Age 9 N	Age 10 N	Age 11 N
ITBS/FY77, FY78 CRT/FY77, FY78				
(Level I)	348* (.87,.87)	187	61	33
(Level II)	78	273* (.92,.95)	551*(.94,.92)	670*(.92,.94)
ITBS/FY77, FY78 CRT/FY77-Level I and FY78-Level II	455	387	256	117
CRT/FY77 - Level II and FY78- Level I	16	28	21	32
ITBS/none or unmatched CRT (all combinations)	197	96	107	91

CRT pre-and posttest reliabilities (Cronbach's α) respectively are given in parenthesis.

TABLE 2A

PEARSON CORRELATIONS BETWEEN NORMED-REFERENCED (ITBS) TESTS AND CRITERION-REFERENCED (CRT) TESTS[a]

	Pretest			Posttest		
Pretest	ITBS GE	ITBS CSS	CRT	ITBS GE	ITBS CSS	CRT
ITBS-GE						
Age 8	1.00	.88	.38	.33	.33	.21
Age 9	1.00	.99	.52	.45	.44	.28
Age 10	1.00	.99	.46	.58	.58	.33
Age 11	1.00	.68	.46	.68	.67	.22
ITBS-CSS						
Age 8	.88	1.00	.36	.28	.31	.18
Age 9	.99	1.00	.53	.44	.44	.27
Age 10	.99	1.00	.46	.58	.59	.33
Age 11	.68	1.00	.32	.49	.38	.18
CRT						
Age 8	.38	.36	1.00	.35	.36	.41
Age 9	.52	.53	1.00	.54	.54	.36
Age 10	.46	.46	1.00	.36	.36	.34
Age 11	.45	.32	1.00	.47	.46	.26
Posttest						
ITBS-GE						
Age 8	.33	.28	.35	1.00	.99	.56
Age 9	.45	.44	.54	1.00	.99	.33
Age 10	.58	.58	.36	1.00	.99	.40
Age 11	.68	.49	.47	1.00	.98	.23
ITBS-CSS						
Age 8	.33	.31	.36	.99	1.00	.56
Age 9	.44	.44	.54	.99	1.00	.33
Age 10	.58	.59	.36	.99	1.00	.39
Age 11	.67	.38	.46	.98	1.00	.23
CRT						
Age 8	.21	.18	.41	.56	.56	1.00
Age 9	.28	.27	.36	.33	.33	1.00
Age 10	.33	.33	.34	.40	.39	1.00
Age 11	.22	.18	.26	.23	.23	1.00

a correlations were all significant $p < .01$.

TABLE 2B

DESCRIPTIVE STATISTICS FOR NORM-REFERENCED (ITBS) AND CRITERION-REFERENCED (CRT) TESTS

	N	Mean	Median	Mode	Variance	Skewness[a]	Kurtosis[b]
Pretest							
ITBS-GE[c]							
Age 8	348	17.05	16.80	17.00	34.99	.678	.548
9	273	28.96	27.94	27.00	56.30	.572	.687
10	551	31.10	30.59	31.00	63.28	.337	.153
11	670	33.76	32.46	31.00	72.44	.745	2.196
ITBS-CSS[d]							
Age 8	348	229.21	228.78	229.00	137.98	.703	.518
9	273	240.68	239.03	237.00	187.66	.444	.946
10	551	236.72	236.56	237.00	168.78	.319	.121
11	670	233.32	231.50	230.00	128.48	.715	2.079
CRT							
Age 8	348	28.49	27.30	27.00	74.48	.262	-.144
9	273	62.50	64.87	65.00	125.08	-.873	.302
10	551	64.80	68.48	75.00	157.66	-1.258	1.114
11	670	69.39	72.01	74.00	93.63	-1.862	4.539
Posttest							
ITBS-GE[c]							
Age 8	348	22.27	21.75	19.00	48.93	.360	-.279
9	273	33.34	33.34	33.00	70.30	.140	-.603
10	551	37.32	37.67	38.00	82.57	.228	.484
11	670	39.69	39.57	40.00	97.98	.439	.978
ITBS-CSS[d]							
Age 8	348	228.76	227.83	223.00	158.72	.359	-.294
9	273	240.48	240.34	240.00	182.01	.135	-.616
10	551	238.01	238.64	239.00	148.33	.202	.411
11	670	234.92	234.62	235.00	142.74	1.046	4.180
CRT							
Age 8	348	37.17	38.93	41.00	74.72	-.761	.076
9	273	68.81	72.24	77.00	138.74	-2.093	4.593
10	551	71.70	74.75	76.00	81.90	-2.172	5.949
11	670	71.49	75.20	78.00	110.53	-2.125	4.743

[a] Positive values indicate a positive skew, negative values a negative skew, zero is the value for a symetrical distribution.

[b] Positive values indicate a "peaked" distribution, negative values indicate a "flattened" distribution, zero is the value for a normal distribution.

[c] ITBS standard GE scores, multiplied by 10.

[d] Chicago standard scores.

REFERENCES

Angoff, William H., Scales, norms, and equivalent scores. In R. L. Thorndike, (ed) Educational Measurement, 2nd edition, Washington, D.C.: American Council on Education, 1971.

Angoff, William H., Can useful general-purpose equivalency tables be prepared for different college admissions tests? In Anne Anastasi (ed) Testing Problems in Perspective, Washington, D.C.: American Council on Education, 1966.

Fisher, R. A. and Yates, F. Statistical tables for biological, agricultural and medical research (4th ed). London: Oliver & Boyd, 1953.

Kearns, Donald R., ESEA Title I Evaluation model equivalence and comparability with real data, Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, 1978.

Linn, Robert L., The validity of inferences based on the proposed Title I evaluation models. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, 1978.

Magnusson, D. Test Theory. Reading Massachusetts: Addison-Wesley, 1966.

Rice, W., Norms for the Chicago Title I Evaluation, Chicago Board of Education, March 1977.

Storlie, T.R., Rice, W., Johnson, M.L., and Crowe, M.B. Local norms in a large urban setting for evaluating Title I programs with Model A 1. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Tallmadge, G.K., and Wood, C.T., User's Guide, ESEA Title I Evaluation and Reporting System, 1978 (Revised), RMC Research Corporation.

Wick, J. W., and Ward, F., Testing students at functioning reading level. A two-year report from Chicago. Chicago Board of Education.